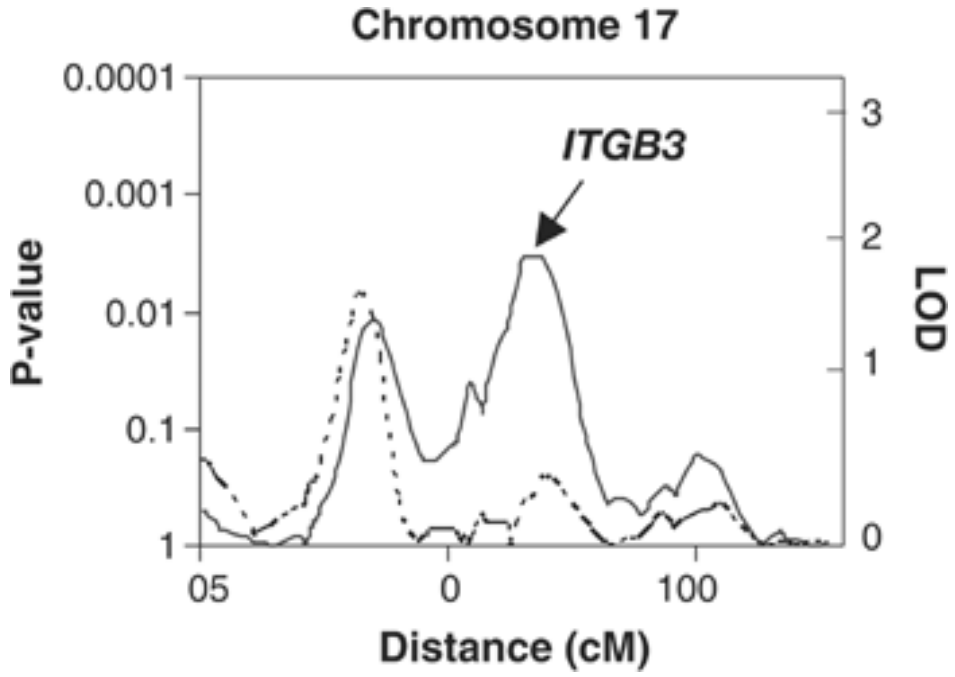
Chapter 2 Class Notes – Words and Probability

Medical/Genetics Illustration – reference Bojesen et al (2003), "Integrin β_3 Leu33Pro Homozygosity and Risk of Cancer", *J. NCI*.



Women only 2 x 2 table:

		Outcome Status		
		With Cancer	Without Cancer	Total
Stratification (Type)	Non-carriers	501 (14.4%)	2983	3484
	Homozygotes	29 (21.5%)	106	135
	Total	530	3089	3619

The variables here are

X = stratification (either non-carrier or homozygote group)

Y = cancer status at end of study.

Both variables here are nominal (and therefore qualitative).

For a review of basic probability (STAT-335), see Chap3 notes at
http://webpages.math.luc.edu/~tobrien/courses/new335/course-homepage.html

Given a single-strand sequence such as (another is on p.37):

ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTGCC
CCTGGAGGGTGGCCCCACCGGCCGAGACAGCGAGCATATGCAGGAAGCGGCAGGAATAAGGAAAAGCAGC
CTCCTGACTTTCCTCGCTTGGTGGTTTGAGTGGACCTCCCAGGCCAGTGCCGGGCCCCTCATAGGAGAGG
AAGCTCGGGAGGTGGCCAGGCGGCAGGAAGGCGCACCCCCCCAGCAATCCGCGCGCCGGGACAGAATGCC
CTGCAGGAACTTCTTCTGGAAGACCTTCTCCTCCTGCAAATAAAACCTCACCCATGAATGCTCACGCAAG
TTTAATTACAGACCTGAA

Summarize/analyze using statistics: is this a coding region? Does this segment differ from other regions? What is its function?

Words – of length:

- **1 are 1-tuples (e.g., nucleotides or purines/pyrimidines)**
- **2 are 2-tuples (e.g., di-nucleotides)**
- **3 are 3-tuples or triples (e.g., codons)**
- **etc.**

DNA sequences from different sources/regions of a genome may be distinguished from each other by their k-tuple content.

2.2. Base Composition (k=1): Consider for a moment the duplex DNA and residues A, C, G and T. Notice that fr(A) = fr(T), fr(C) = fr(G); since fr(A) + fr(C) + fr(G) + fr(T) = 1, it follows that fr(A+T) = 1- fr(C+G). So, you only need to know fr(C+G). These are given for different organisms on p.39: these percentages range from 31.6% for *Mycoplasma genitalium* to 66.4% for *Pseudomonas aeruginosa* PAO1 (bacteria). Asymmetries can be detected especially on the leading strand using the $GC\ skew = \frac{\#G - \#C}{\#G + \#C}$

2.3. Introduction to Probability (review)

A discrete random variable (RV) X takes on values $x_1, x_2, \ldots, x_J$ with respective probabilities $p_1, p_2, \ldots, p_J$ and such that $p_1 + p_2 + \cdots + p_J = 1$. Associating the probabilities with the realized values of X either in a table, formula or graph is the *probability mass function* of X. An example could be for nucleotides with $p_A = 0.25, p_C = 0.25, p_G = 0.25, p_T = 0.25$; we can then derive a second RV which is only counts A's: it is 1 if the nucleotide is A and 0 otherwise; this is called a Bernoulli RV.

Now, consider a series of Bernoulli RVs X_1, X_2, …, X_n corresponding to positions 1, 2 … n; again, each of these is 1 if the respective nucleotide is A and 0 otherwise. For each of these, $P(X_i = 1) = P(L_i = A) = p_A$; also, $P(X_i = 0) = P(L_i = C \text{ or } G \text{ or } T) = p_C + p_G + p_T = 1 - p_A$. Then $N = X_1 + X_2 + \cdots X_n$ is a RV that counts the number of A's in the sequence of length n. Next, we'll make a big assumption (for the moment): that what occurs at one position is independent of what occurs elsewhere (see STAT-335 notes and p.42 §2.3.2). In the case of independence, probabilities of intersections (AND) amount to simply multiplying the unconditional probabilities.

2.3.3. Expected Values and Variances: the expected value (EV) of RV X is just its mean: $E(X) = \mu = x_1 p_1 + x_2 p_2 + \cdots x_J p_J$. Thus, for the above $E(X_i) = p_A$. Also, since $E(X_1 + X_2 + \cdots X_n) = E(X_1) + E(X_2) + \cdots E(X_n)$, $E(N) = np_A = np$ (dropping the subscript for convenience). Next, variance is defined as $Var(X) = E[(X - \mu)^2] = \sum_{k=1}^{J} (x_k - \mu)^2 p_k$; the short-cut

formula is then $Var(X) = E(X^2) - \mu^2 = \sum_{k=1}^{J} x_k^2 p_k - \mu^2$. So, $Var(X_i) = p_A - p_A^2 = p_A(1 - p_A)$ by the short-cut formula. A final snippet: since we're assuming the X_i's are independent and since the variance of a sum of independent RVs is the sum of the individual variances, $Var(N) = nVar(X_i) = np_A(1 - p_A) = np(1 - p)$ (again dropping the subscript).

2.3.4. The Binomial Distribution: by now, you have noticed that $N = X_1 + X_2 + \cdots X_n$ is a Binomial RV with parameters n and $p_A = p$. Hence, $P(N = k) = \binom{n}{k} p^k (1 - p)^{n-k}, k = 0, 1, \ldots, n$.

2.4. Simulating from Probability Distributions: to understand the behavior (e.g. distribution) of a RV such as N, simulation can be very helpful. To do so, we'd generate a large number (m) of N: $N_1, N_2, \ldots, N_m$. We could then calculate the sample mean ($\bar{N} = \frac{1}{m}\sum_{i=1}^{m} N_i$), sample variance ($s^2 = \frac{1}{m-1}\sum_{i=1}^{m}(N_i - \bar{N})^2$), and histogram to get an idea of $E(N)$, $Var(N)$, and the distribution of N. To do this, pseudo-random number generators are used; these generators simulate uniform RVs on the interval $(0, 1)$, and then convert. To illustrate, it's easy to simulate from $p_A = 0.25, p_C = 0.25, p_G = 0.25, p_T = 0.25$ distribution (a) for A versus rest, and (b) for all four nucleotides using Minitab and R.

In R, we use the "sample" command:

```
pi<-c(0.25,0.75)
x<-c(1,0)
seq1<-sample(x,10000,replace=T,pi)
seq1[1:30]
```

For the nucleotide problem:

```
pi<-c(0.25,0.25,0.25,0.25)
x<-c(1,2,3,4)
seq<-sample(x,10000,replace=T,pi)
seq[1:30]
hist(seq)
```

To simulate from the Binomial ($n = 1000, p = 0.25$) distribution we use the "rbinom" command:

```
x<-rbinom(10000,1000,0.25)
x[1:10]
mean(x)
sd(x)^2
hist(x,xlab="number of successes",main="Binomial Simulation")
```

Now, suppose in a sequence of length $n = 1000$, we observe $y = 280$ A's, and we want to know the associated p-value in a test where the alternative hypothesis is $H_1: p_A > 0.25$; then we wish to find the probability of at least 280 'successes' in this binomial experiment. Ways to find this are (a) direct calculation from equation (2.20) on p.48 ($ans = 0.016437$), (b) approximation using the CLT ($P\left(Z > \frac{279.5-250}{\sqrt{187.5}}\right) = 0.015605$), (c) using Minitab or calculator, or (d) simulation (see p.63 exercise 4).

2.5. Biological Words (k=2): let l_i denote the nucleotide at position i and so $l_i l_{i+1}$ is the dinucleotide starting at the same position (reading of course in the 5' to 3' direction); l_i can take on values A , C , G , T, and $l_i l_{i+1}$ can equal AA , AC , AG , AT , CA , CC ,

CG, CT, GA, GC, GG, GT, TA, TC, TG, TT. "Dinucleotides are important in part because physical parameters associated with them can describe the trajectory of the DNA helix through space (DNA bending), which may have effects on gene expression."

Next, with $P(L_i = r_1) = p_{r_1}$ and $P(L_{i+1} = r_2) = p_{r_2}$, we can test for independence, $P(L_i = r_1, L_{i+1} = r_2) = p_{r_1}p_{r_2}$ for all pairs, by using $X^2 = \frac{(O-E)^2}{E}$ with $E = (n-1)p_{r_1}p_{r_2}$. First, we need to calculate

$$c = \begin{cases} 1 + 2p_{r_1} - 3p_{r_1}^2, & \text{if } r_1 = r_2 \\ 1 + 3p_{r_1}p_{r_2}, & \text{if } r_1 \neq r_2 \end{cases}$$

Then, the relevant test statistic is $\chi_1^2 = X^2/c$, so for example, if $X^2/c > 3.84$ then we conclude that the iid model does not fit for this dinucleotide pair at the 5% level of significance. In practice, frequencies from the data fr(A), fr(C), fr(G), fr(T) are used to estimate the probabilities. Empirical results are given in Table 2.2 on p.50 for *E. coli* and *Mycoplasma genitalium*; clearly, the first 1000 bp for *E. coli* deviate from the *iid* model.

Table 2.2. Observed values of X^2/c for the first 1000 bp of each genome. For *E. coli*, the base frequencies were taken as $(0.25, 0.25, 0.25, 0.25)$, whereas for *M. genitalium* they were (0.45, 0.09, 0.09, 0.37), close to the observed frequencies. Significant values are in bold.

	Observed X^2/c for	
Dinucleotide	*E. coli*	*M. genitalium*
AA	**6.78**	0.15
AC	0.05	1.20
AG	**5.99**	0.18
AT	0.01	0.01
CA	2.64	0.01
CC	0.03	0.39
CG	0.85	**4.70**
CT	**4.70**	1.10
GA	2.15	0.34
GC	**10.04**	1.07
GG	0.01	0.09
GT	1.76	0.61
TA	**5.99**	1.93
TC	**9.06**	2.28
TG	3.63	0.05
TT	1.12	0.13

2.6. Introduction to Markov Chains: As seen by Table 2.2, the *iid* model may not be rich enough, and a richer one may be needed – one approach is to use a Markov chain. If in a nucleotide at position we observe a C, then the nucleotide at the next position may depend upon this information, and we are led to conditional probabilities; often we find that in this case the probability that the next position is G is less than its being an A, C, or T.

2.6.1. Conditional Probability (review): in the introductory notes, we saw that the probability of events A and B (intersection) occurring is $P(A \cap B)$ and the probability of events A or B (union) occurring is $P(A \cup B)$. Then, the conditional probability of event A given that B has occurred is:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Similarly, $P(B|A) = \frac{P(A \cap B)}{P(A)}$. An immediate consequence of this is the multiplication rule: $P(A \cap B) = P(A)P(B|A) = P(B)P(A|B)$. Bayes' Theorem follows directly:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} \quad (*)$$

Also, let $B_1, B_2, \dots, B_k$ for a partition of Ω (so that all B_i are disjoint $B_i \cap B_j$ for $i \neq j$ and exhaustive $B_1 \cup B_2 \cup \dots \cup B_k = \Omega$), then $P(A) = \sum_{i=0}^{k} P(A|B_i)P(B_i)$, so (*) can be rewritten.

2.6.2. The Markov Property: "the sequence $\{X_t, t = 0, 1, \dots\}$ is called a first-order Markov chain if the probability of finding a particular character at position $t + 1$ given the preceding characters at positions $t, t - 1, \dots 0$ is identical to the probability

of observing the character at position $t+1$ given the character state of the immediate preceding position, t." Formally,

$$P(X_{t+1}=j|X_t=i,X_{t-1}=i_{t-1},\dots,X_0=i_0)=P(X_{t+1}=j|X_t=i)$$

Markov chains of order 2 correspond to the case where the conditional probability of the present position depends on the previous 2 positions, and so on. We also consider only Markov chains that are homogeneous: the above property doesn't depend upon t (position along the nucleotide or segment).

Next, let $p_{ij}=P(X_{t+1}=j|X_t=i)$ be the one-step transition probabilities for $i,j\in\mathfrak{X}$; for the nucleotide problem, $\mathfrak{X}=\{\mathtt{A},\mathtt{C},\mathtt{G},\mathtt{T}\}$. For the nucleotide problem, these probabilities are given in the one-step transition matrix:

$$P=\begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

Each row sums to one, so $\sum_j p_{ij}=1,\forall i$; this follows since after each position there must be one character from $\mathfrak{X}$ next in the sequence, and can be written in the nucleotide case as

$$P\vec{1}=\vec{1},\vec{1}=\begin{bmatrix}1\\1\\1\\1\end{bmatrix}$$

Next, we have to indicate how the Markov chain begins; for this, we let the initial probability distribution π have elements

$$\pi_i=P(X_0=i),i\in\mathfrak{X}$$

Also, at time 1, we have $P(X_1 = j) = \sum_i \pi_i p_{ij}$; in matrix terms: $\pi^{(1)} = \pi P$. In the nucleotide case, the right hand side is

$$[\pi_A \quad \pi_C \quad \pi_G \quad \pi_T] \begin{bmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{bmatrix}$$

The result is also a 1×4 matrix (vector), and the first element is

$$P(X_1 = A) = \pi_A p_{AA} + \pi_C p_{CA} + \pi_G p_{GA} + \pi_T p_{TA}$$

On p.53 (text), the authors show that the transition matrix from time t to time $t + 2$ is given by P^2, and so on, so that means that $\pi^{(2)} = \pi^{(1)} P = \pi P^2$.

A *stationary distribution* π for the chain is when $\pi_j = \sum_i \pi_i p_{ij}, \forall j$, so in matrix form: $\pi = \pi P$.

2.6.3. A Markov Chain Simulation: following is the dinucleotide frequencies of *M. genitalium*:

	A	C	G	T
A	0.146	0.052	0.058	0.089
C	0.063	0.029	0.010	0.056
G	0.050	0.030	0.028	0.051
T	0.087	0.047	0.063	0.140

Summing across the rows, we obtain empirical estimates $p_A = 0.345, p_C = 0.158, p_G = 0.159, p_T = 0.337$, and then dividing through (e.g. $p_{AA} = \frac{0.146}{0.345} = 0.423$), we get $P =$

	A	C	G	T
A	0.423	0.151	0.168	0.258
C	0.399	0.184	0.063	0.354
G	0.314	0.189	0.176	0.321
T	0.258	0.138	0.187	0.415

We can also assume that the initial distribution is (as above)

$$\boldsymbol{\pi} = (0.345, 0.158, 0.159, 0.337)$$

Now we can simulate (see Computational Example 2.3 on p.55):

```
markov1 <- function(x,pi,P,n) {
   mg <- rep(0,n)
   mg[1] <- sample(x,1,replace=TRUE,pi)
   for(k in 1:(n-1)) {
      mg[k+1]<-sample(x,1,replace=T,P[mg[k],])  }
   return(mg) }

x<-c(1:4)
pi<-c(0.342,0.158,0.158,0.342)
P<-matrix(scan(),ncol=4,nrow=4,byrow=T)
0.423 0.151 0.168 0.258 0.399 0.184 0.063 0.354 0.314
0.189 0.176 0.321 0.258 0.138 0.187 0.415
tmp<-markov1(x,pi,P,50000)
length(tmp[tmp[]==1])/length(tmp)
[1] 0.3441
length(tmp[tmp[]==2])/length(tmp)
[1] 0.15882
```

```
length(tmp[tmp[]==3])/length(tmp)
[1] 0.15936
length(tmp[tmp[]==4])/length(tmp)
[1] 0.33772
```

Also, we used the above to simulation the *dinucleotide* distribution (not shown); since it matches the original P, we conclude as in the text (p.57) that "the Markov model provides a good probabilistic description of the data for *M. genitalium* DNA."

2.7. Biological Words with k=3 (Codons): We'll discuss two measures here relating codons to amino acids with an eye to distinguishing coding from non-coding genetic regions.

First, we'll compare expressed codon frequencies to those which would be expressed under independence. To illustrate using *E. coli*, from Table 2.1 (p.39), $\widehat{P}(A) = \widehat{P}(T) = 0.246$, $\widehat{P}(C) = \widehat{P}(G) = 0.254$, so under independence, $\widehat{P}(TTT) = 0.246 \times 0.246 \times 0.246 = 0.01489$ and $\widehat{P}(TTC) = 0.246 \times 0.246 \times 0.254 = 0.01537$. These are the only two codons coding for amino acid Phe (phenylaline), so under independence we expect the TTT proportion to be roughly $\frac{0.01489}{0.01489+0.01537} = 0.492$. Similar results are given in the "Predicted" column in Table 2.3, and these results are compared with the empirical results for two gene classes (I and II). Ménigue *et al* then noted when there were big differences and investigated and found that "Class II genes were largely those such as ribosomal proteins or translation factors – genes expressed at high levels – whereas Class I genes were mostly those that are expressed at moderate levels."

Table 2.3. Comparison of predicted and observed triplet frequencies in coding sequences for a subset of genes and codons from *E. coli*. Figures in parentheses below each gene class show the number of genes in that class. Data were taken from Médigue et al. (1991).

	Codon	Predicted	Observed Gene Class I (502)	Observed Gene Class II (191)
Phe	TTT	0.493	0.551	0.291
	TTC	0.507	0.449	0.709
Ala	GCT	0.246	0.145	0.275
	GCC	0.254	0.276	0.164
	GCA	0.246	0.196	0.240
	GCG	0.254	0.382	0.323
Asn	AAT	0.493	0.409	0.172
	AAC	0.507	0.591	0.828

Another measure is the CAI (codon adaptation index):

$$CAI = \left[\prod_{k=1}^{L} \frac{p_k}{q_k}\right]^{1/L}$$

It "is the geometric mean of the ratios of the probabilities for the codons actually used to the probability of the codons most frequently used in highly expressed genes." It is illustrated on p.59 again using *E. coli*. "In E. coli, a sample of 500 protein-coding genes displayed CAI values in the range from 0.2 to 0.85." Further, "there is a correlation between the CAI and mRNA levels. In other words, the CAI for a gene sequence in genomic DNA provides a first approximation of its expression level: if the CAI is relatively large, then we would predict that the expression level is also large."

M	A	L	T	K	A	E	M	S	E	Y	L	F	...
ATG	GCG	CTT	ACA	AAA	GCT	GAA	ATG	TCA	GAA	TAT	CTG	TTT	...
1.000	0.469 0.057 0.275 0.199	0.018 0.018 0.038 0.033 0.007 0.888	0.451 0.468 0.035 0.046	0.798 0.202	0.469 0.057 0.275 0.199	0.794 0.206	1.000	0.428 0.319 0.033 0.007 0.037 0.176	0.794 0.206	0.193 0.807	0.018 0.018 0.038 0.033 0.007 0.888	0.228 0.772	
ATG	GCT GCC GCA GCG	TTA TTG CTT CTC CTA CTG	ACT ACC ACA ACG	AAA AAG	GCT GCC GCA GCG	GAA GAG	ATG	TCT TCC TCA TCG AGT AGC	GAA GAG	TAT TAC	TTA TTG CTT CTC CTA CTG	TTT TTC	

Fig. 2.2. Example of codon usage patterns in *E. coli* for computation of the codon adaptation index of a gene. The probability for the most frequently used codon in highly expressed genes is underlined.